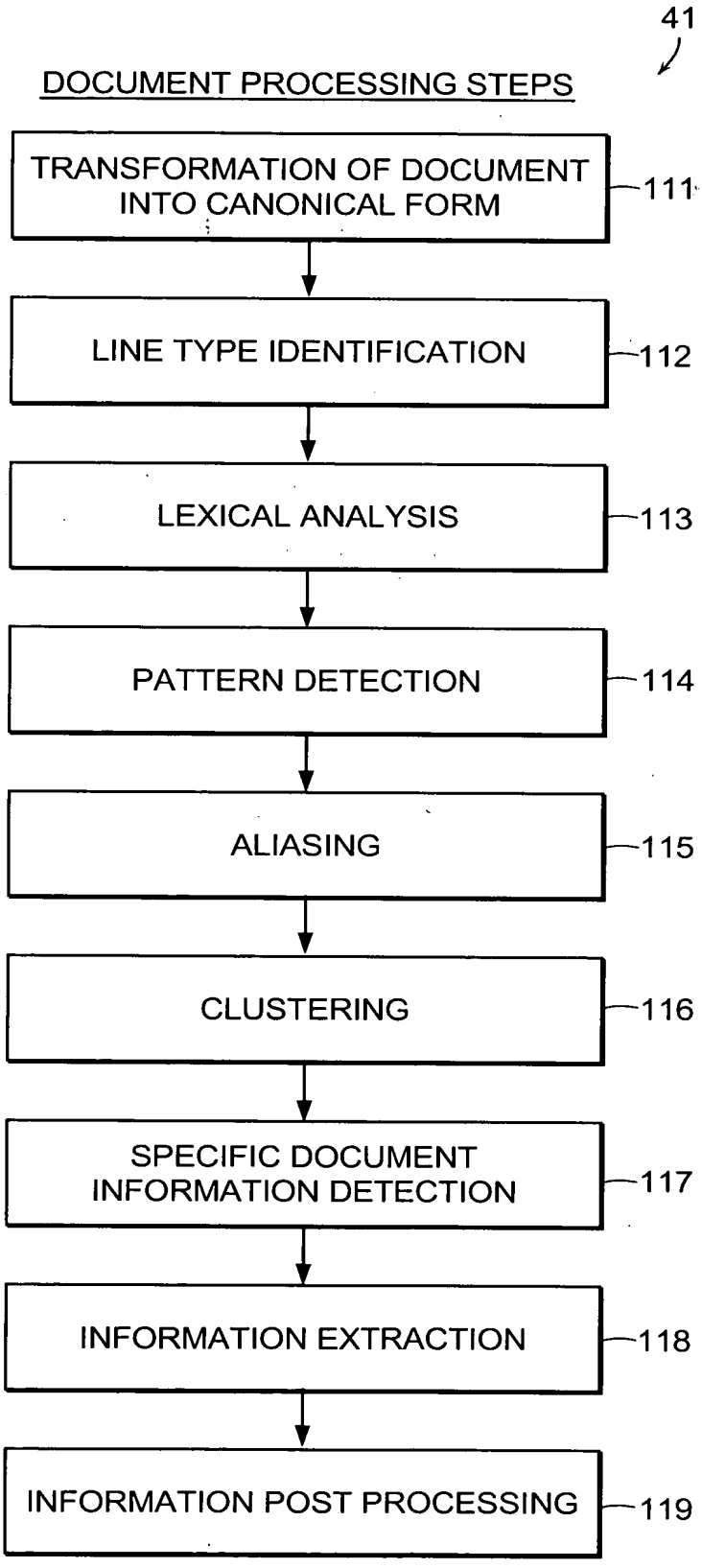
41
DOCUMENT PROCESSING STEPS
TRANSFORMATION OF DOCUMENT INTO CANONICAL FORM
111
LINE TYPE IDENTIFICATION
112
LEXICAL ANALYSIS
113
PATTERN DETECTION
114
ALIASING
115
CLUSTERING
116
SPECIFIC DOCUMENT INFORMATION DETECTION
117
INFORMATION EXTRACTION
118
INFORMATION POST PROCESSING
119

FIG. 1

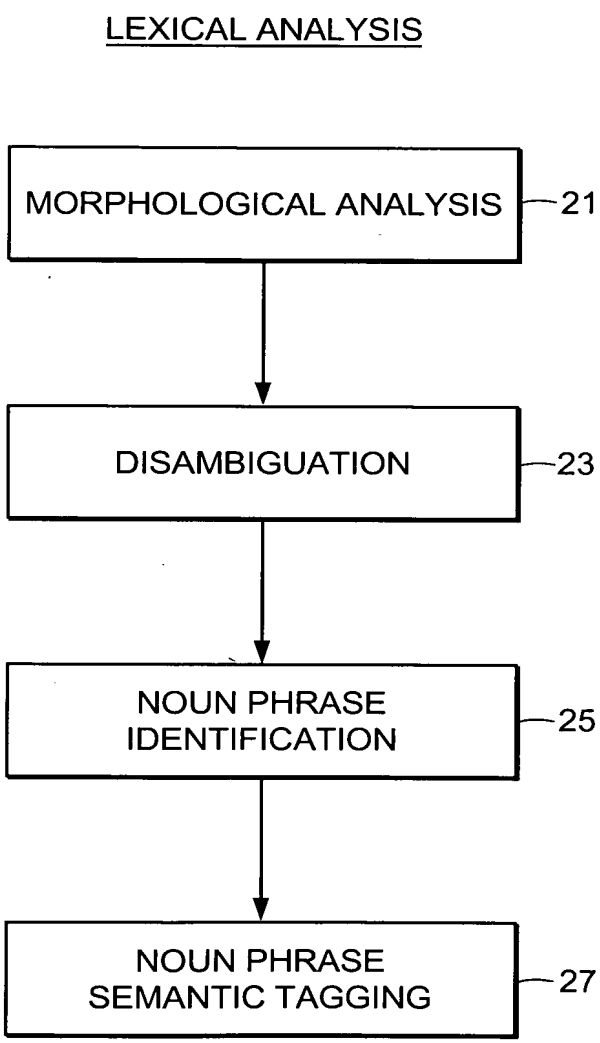
LEXICAL ANALYSIS
113
MORPHOLOGICAL ANALYSIS
21
DISAMBIGUATION
23
NOUN PHRASE IDENTIFICATION
25
NOUN PHRASE SEMANTIC TAGGING
27

FIG. 2

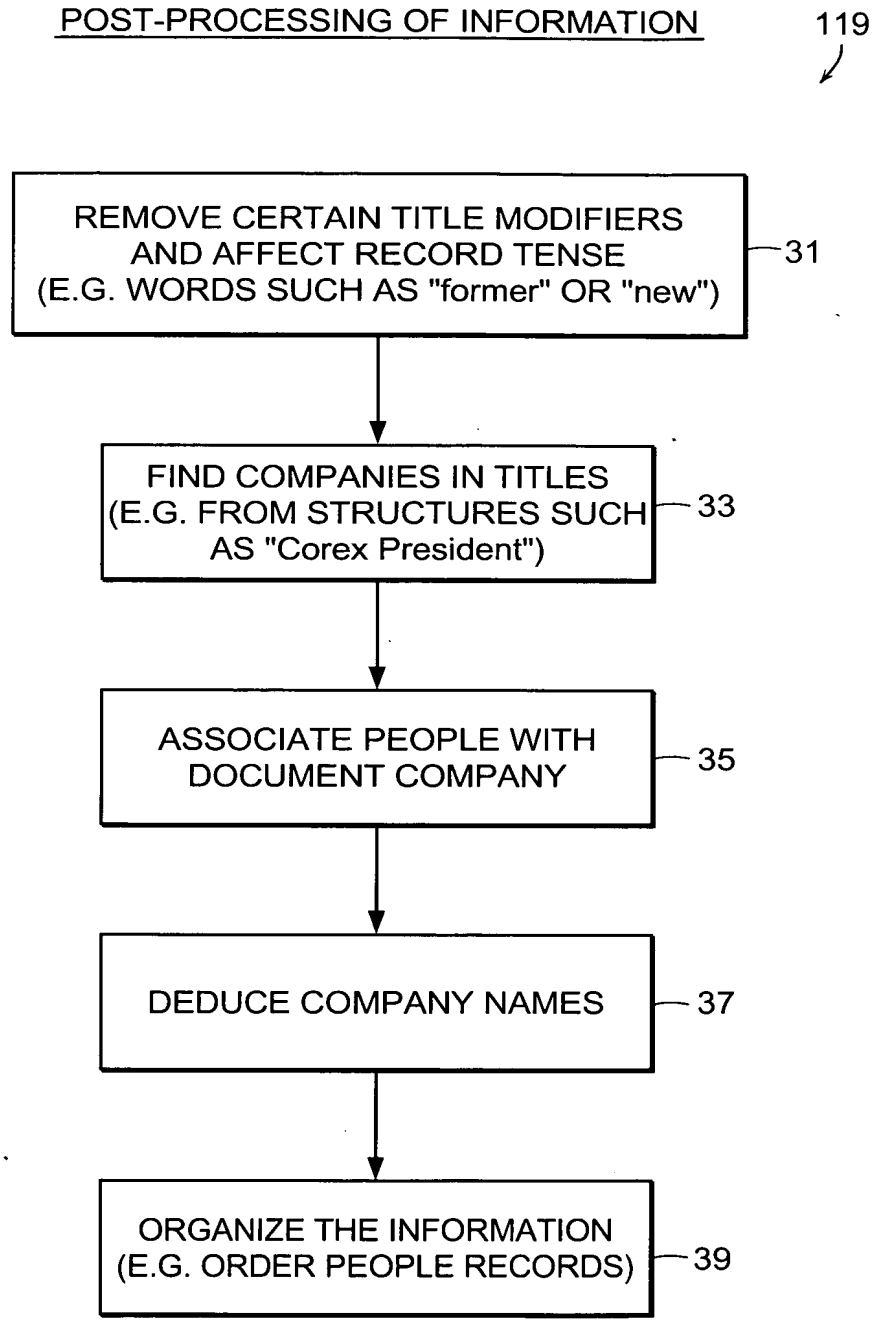
POST-PROCESSING OF INFORMATION
119
REMOVE CERTAIN TITLE MODIFIERS AND AFFECT RECORD TENSE (E.G. WORDS SUCH AS "former" OR "new")
31
FIND COMPANIES IN TITLES (E.G. FROM STRUCTURES SUCH AS "Corex President")
33
ASSOCIATE PEOPLE WITH DOCUMENT COMPANY
35
DEDUCE COMPANY NAMES
37
ORGANIZE THE INFORMATION (E.G. ORDER PEOPLE RECORDS)
39

FIG. 3

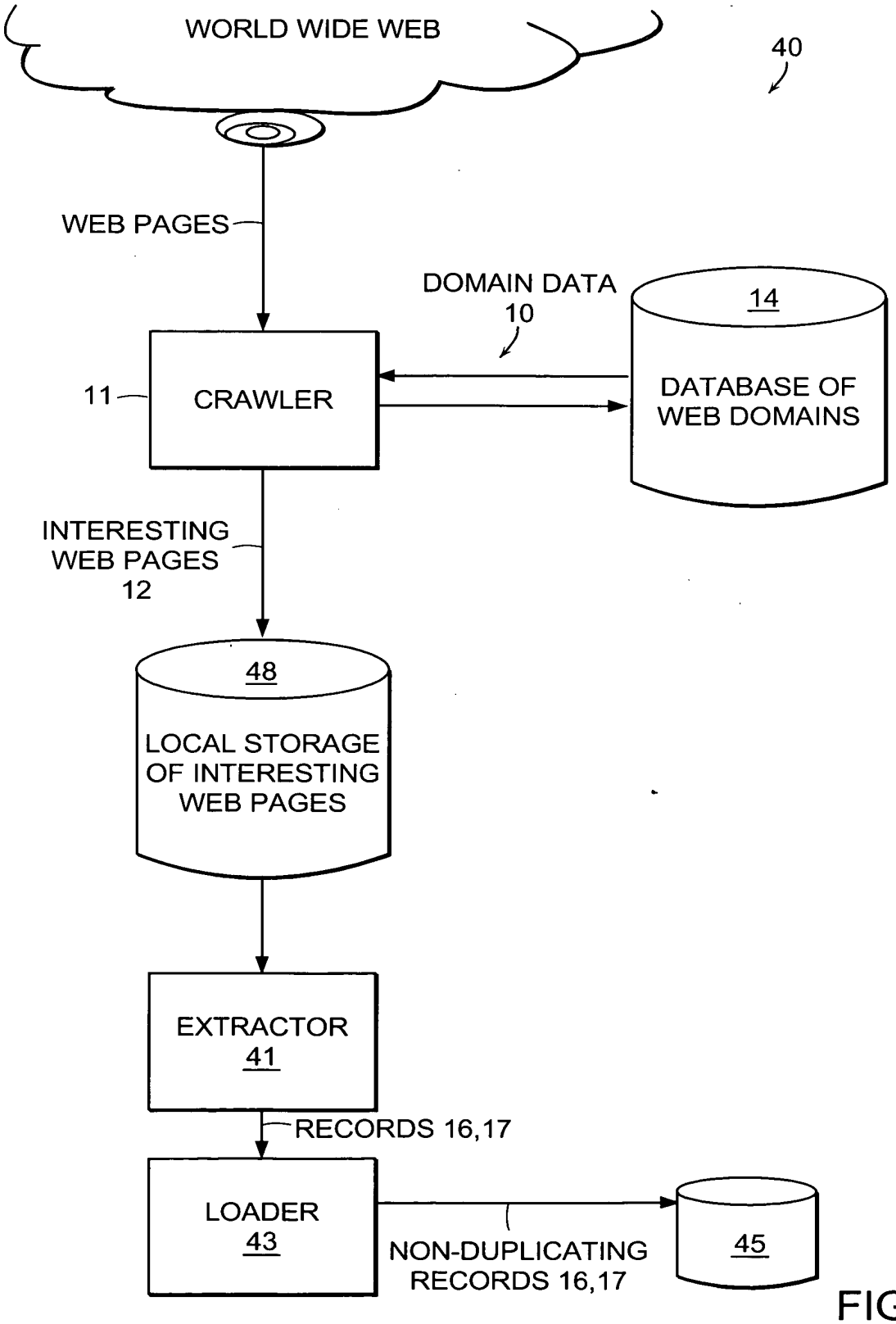
WORLD WIDE WEB
40
WEB PAGES
DOMAIN DATA
10
14
11
CRAWLER
DATABASE OF
WEB DOMAINS
INTERESTING
WEB PAGES
12
48
LOCAL STORAGE
OF INTERESTING
WEB PAGES
EXTRACTOR
41
RECORDS 16,17
LOADER
43
NON-DUPLICATING
RECORDS 16,17
45

FIG. 4

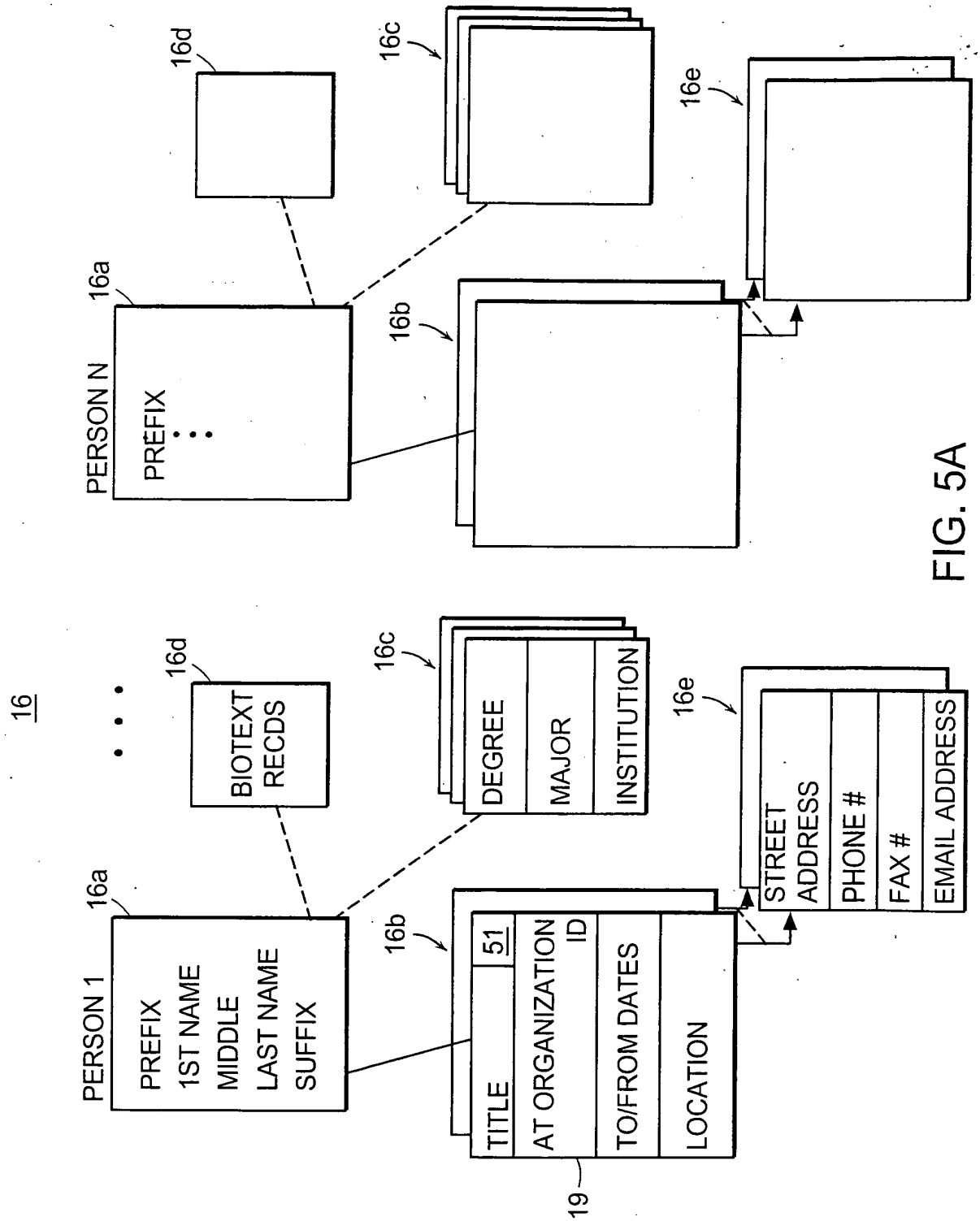
16
PERSON 1
PREFIX
1ST NAME
MIDDLE
LAST NAME
SUFFIX
16a
BIOTEXT RECDS
16d
DEGREE
MAJOR
INSTITUTION
16c
TITLE
51
AT ORGANIZATION
ID
TO/FROM DATES
LOCATION
16b
19
STREET ADDRESS
PHONE #
FAX #
EMAIL ADDRESS
16e
PERSON N
PREFIX
16a
16d
16c
16b
16e

FIG. 5A

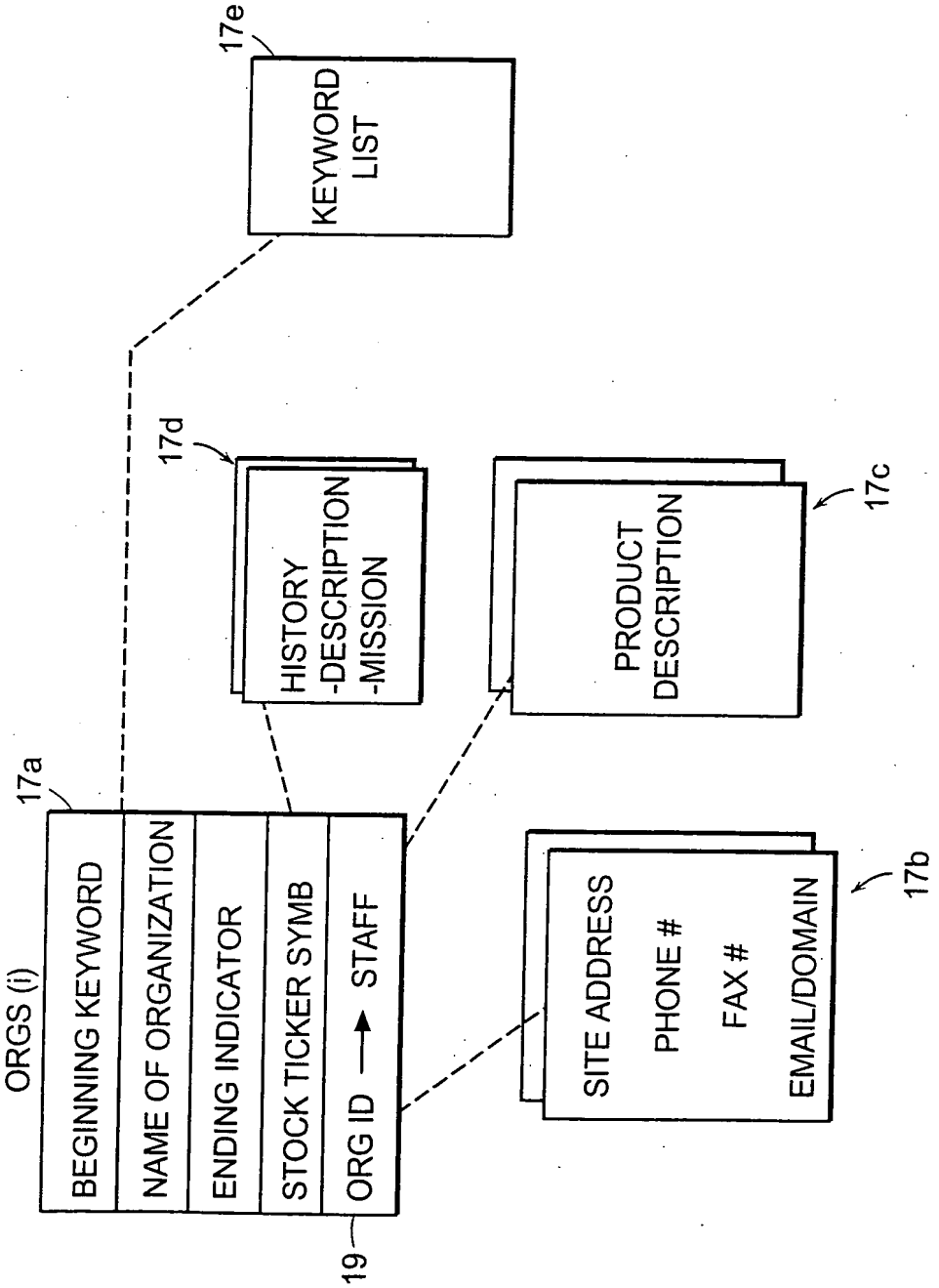
17e
KEYWORD LIST
17d
HISTORY -DESCRIPTION -MISSION
17c
PRODUCT DESCRIPTION
17a
ORGS (i)
BEGINNING KEYWORD
NAME OF ORGANIZATION
ENDING INDICATOR
STOCK TICKER SYMB
ORG ID → STAFF
19
SITE ADDRESS
PHONE #
FAX #
EMAIL/DOMAIN
17b

FIG. 5B